企业大数据技术及应用

Enterprise Big Data: Technology & Applications

塔布 / 数说故事创始人
徐亚波博士

CONTENTS
目 录

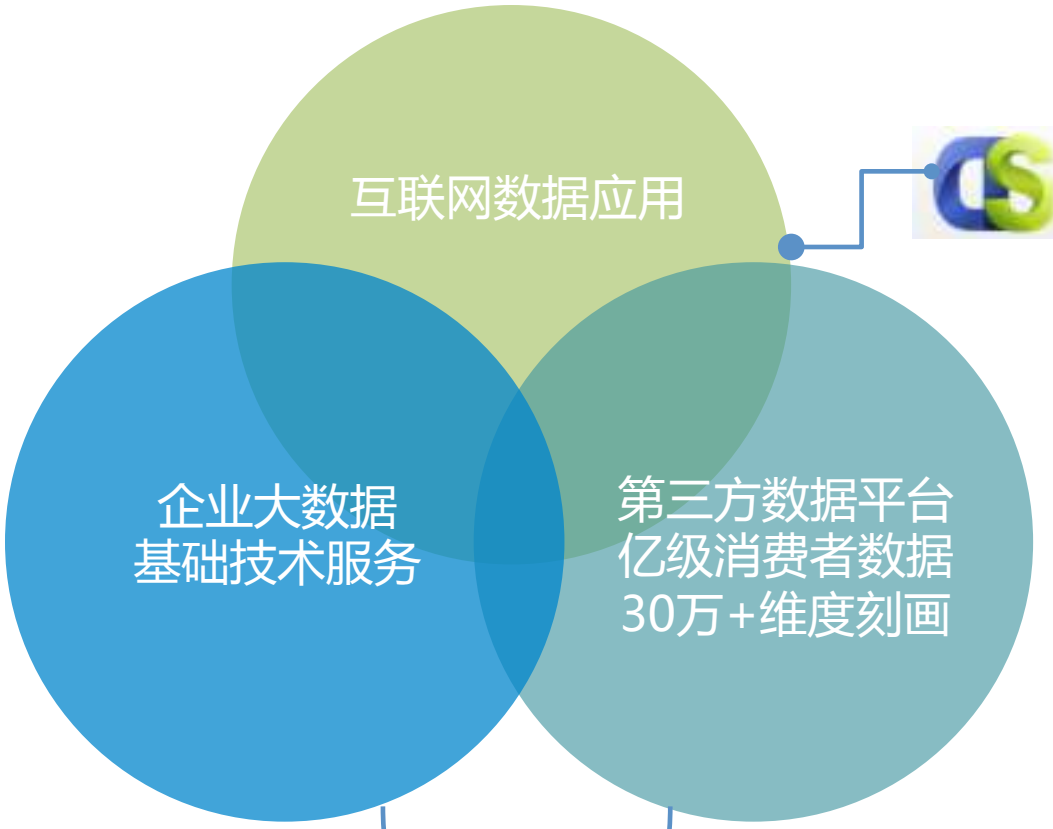
01 关于塔布数说

- 塔布数说简史
- 塔布数说定位

塔布数说简史

Academia		Industry	Data
DM Models & Algo	2001	摘要式答案引擎	100M~1G 结构化数据
Search & Mining	2007	新闻聚合引擎	10G~ 网页数据
Cloud Computing	2009	购物搜索	1T+ 新闻数据
NLP & Text Mining			1T+ 商品/评论数据
	2010		10T~ Social数据
Data Infrastructure &			日更新1G+
Data Integration &	2012	微博分析工具	
Semantic Analysis &	2013	社会化媒体数据平台	
User Modeling &			企业内部+外部数据
Recommendation	2014	数说故事(DATASTORY)	100T ~ 1P
	2015	大数据应用	结构化+非结构化
		塔布(DATATUB):	日更新 1T+
Big Data Services	2016	大数据管理平台	日更新 100T+
		→企业大数据整体服务提供商	

塔布数说定位：企业大数据整体解决方案提供商



02

企业大数据现状与趋势

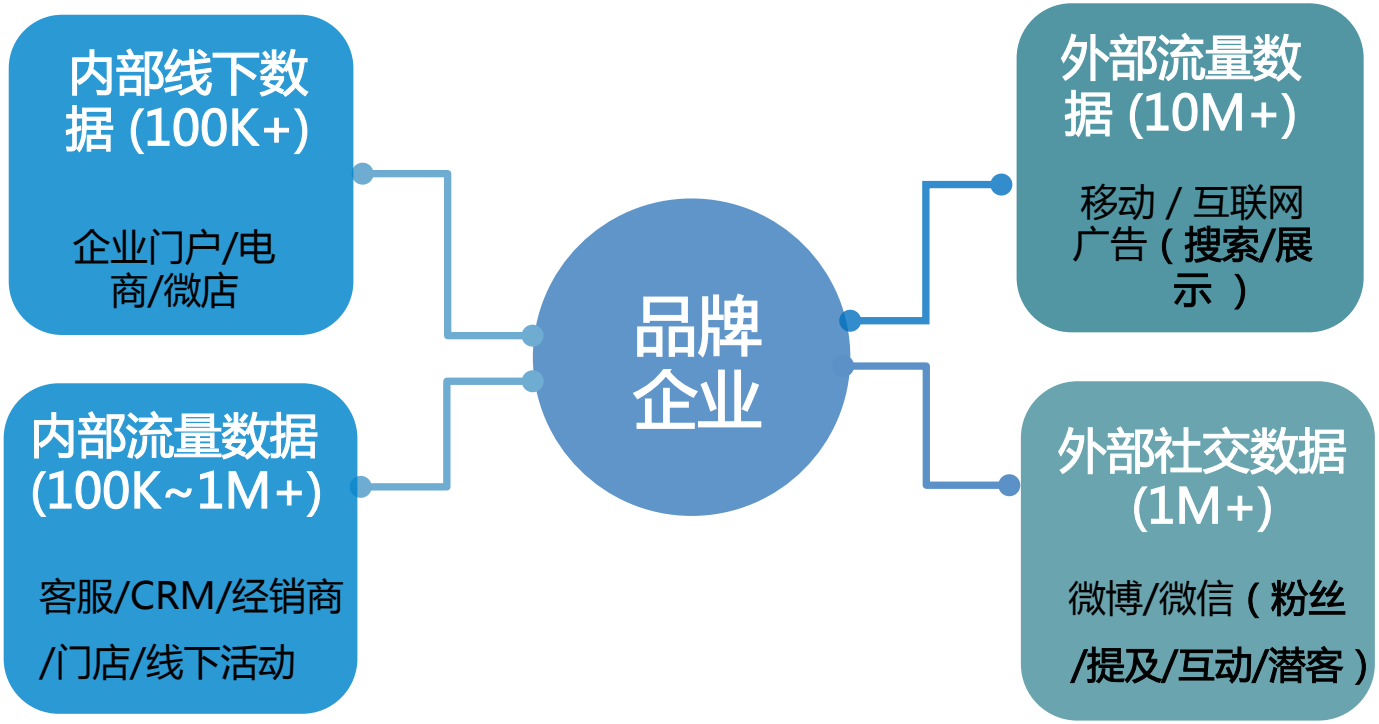
- 为什么企业会有大数据？
- 企业大数据有多少？
- 行业大数据在什么阶段？
- 企业大数据需求阶段

为什么会有企业大数据?

社交 / 移动 / 电商 / 数字广告

企业→消费者: 距离缩短, 接触面扩大, 核心环节数字化

企业有多少数据？


内部线下数据 (100K+)
企业门户/电商/微店
内部流量数据 (100K~1M+)
客服/CRM/经销商/门店/线下活动
品牌企业
外部流量数据 (10M+)
移动 / 互联网广告（搜索/展示）
外部社交数据 (1M+)
微博/微信（粉丝/提及/互动/潜客）

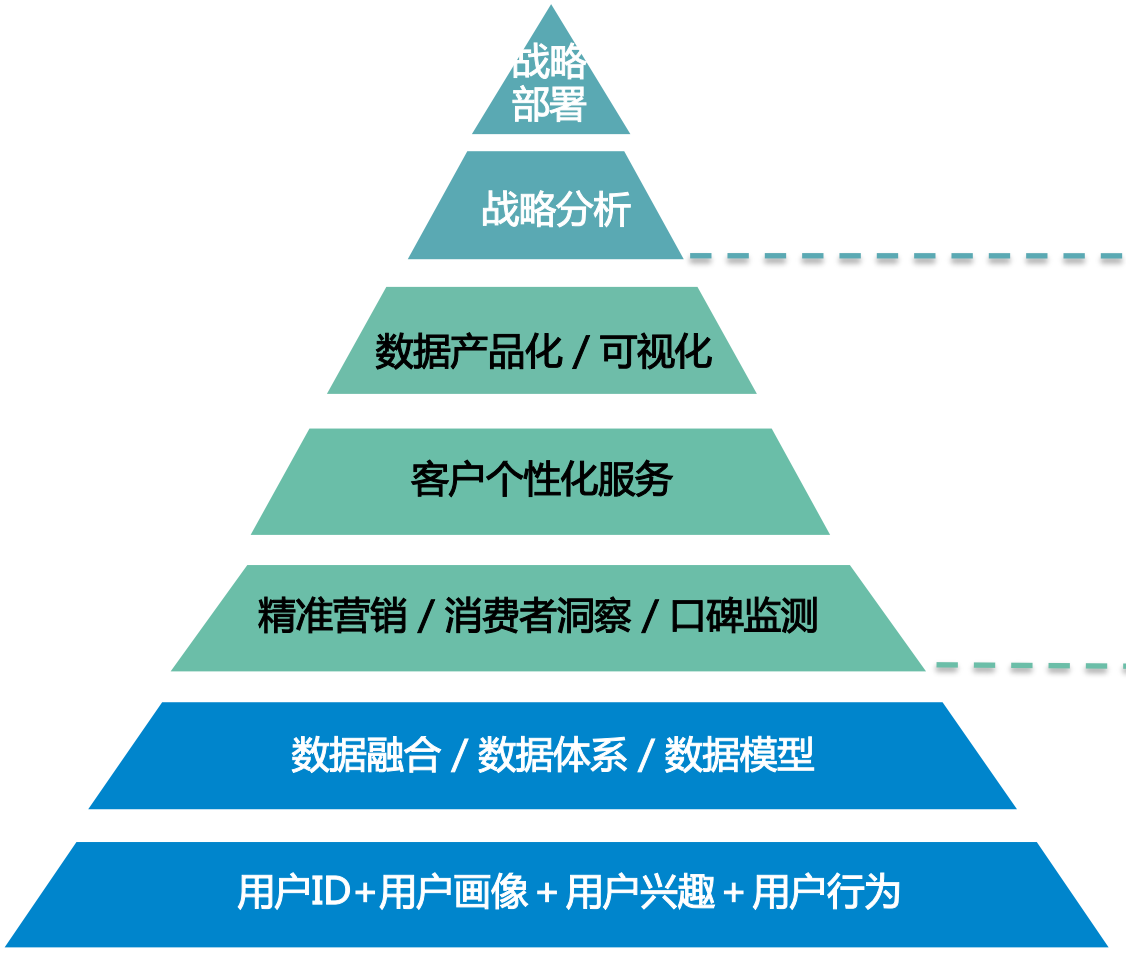
行业大数据在什么阶段？

大数据发展的3个要素

03 大数据核心技术

- 大数据应用体系
- 大数据产品体系
- 大数据平台架构
- 大数据核心技术要素

企业大数据应用金字塔



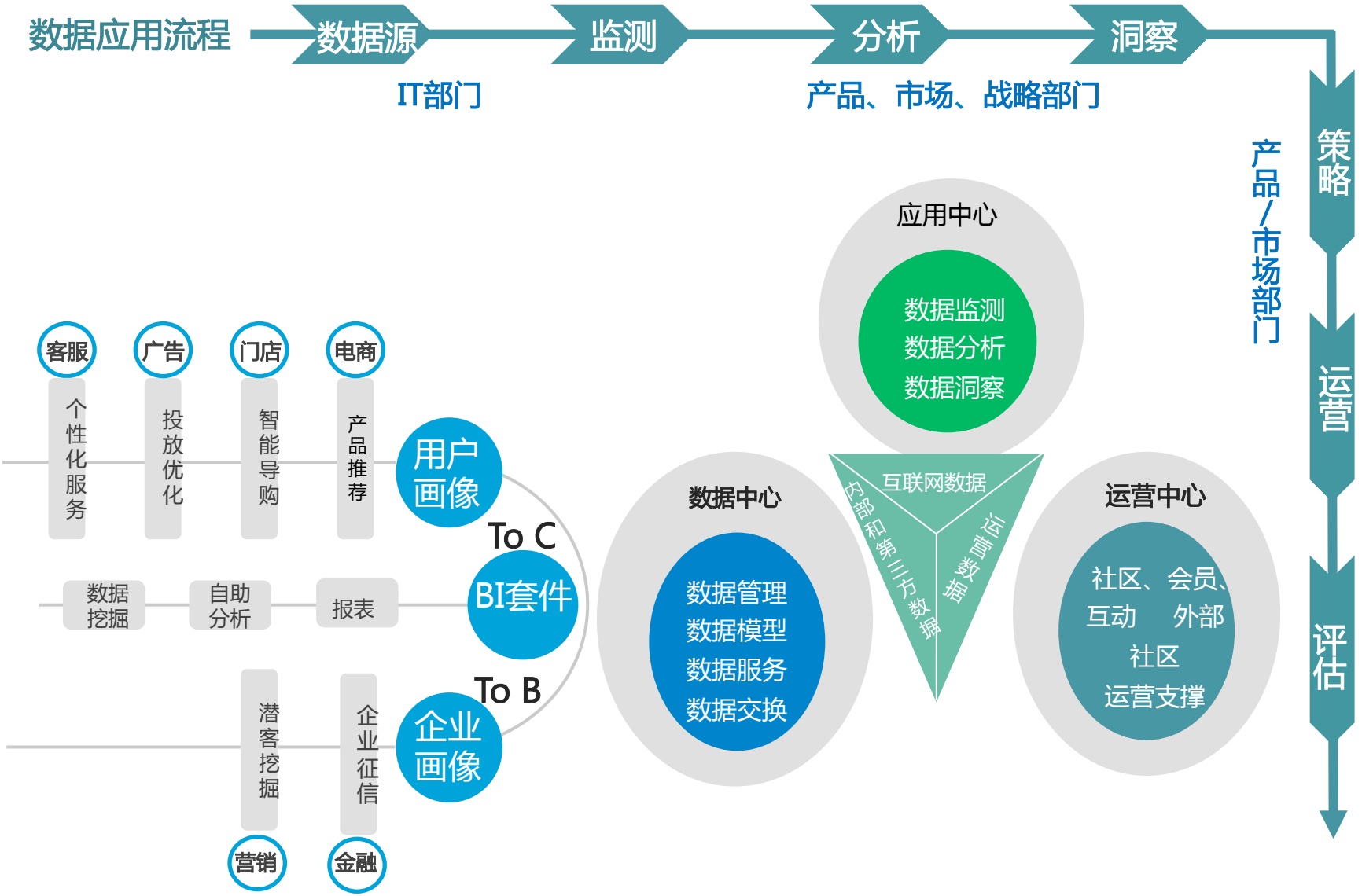
- 战略分析与部署
- 业务经营分析

- 精细化运营和营销
- 用户/客户体验优化
- 数据辅助市场传播

- 大数据基础平台

Ref: http://www.199it.com/archives/235234.html

企业大数据产品体系


数据应用流程
数据源
监测
分析
洞察
IT部门
产品、市场、战略部门
策略
运营
评估
产品/市场部门
应用中心
数据监测
数据分析
数据洞察
数据中心
数据管理
数据模型
数据服务
数据交换
互联网数据
内部和第三方数据
运营数据
运营中心
社区、会员、互动
外部社区
运营支撑
客服
广告
门店
电商
个性化服务
投放优化
智能导购
产品推荐
用户画像
To C
BI套件
数据挖掘
自助分析
报表
To B
企业画像
潜客挖掘
企业征信
营销
金融

高大上的大数据平台架构

数据运营

咨询分析 | 营销策划 | 业务优化 | 产品运营

数据驱动运营实现价值变现

数据 + 运营 → $$

数据应用

市场研究 | 精准营销 | 业务优化 | 个性化服务

数据驱动应用产生价值

数据→洞察→服务质量+营销效果 + 创新应用

数据平台服务

Data as a Service (DaaS)

大数据管理平台(DMP)

把分散的数据作为资产来统一管理
通过数据的融合和分析使数据增值
把数据能力作为服务提供

计算平台服务

Hadoop as a Service (HaaS)

Hadoop统一资源池

强大的统一数据存储、资源调度、资源隔离，
为各种数据服务提供：数据集成、处理、查询能力

数据资源接入

企业内部数据 | 互联网数据

全面的数据源

获取互联网上的数据源，进一步丰富自身数据

HaaS提供统一计算和存储平台

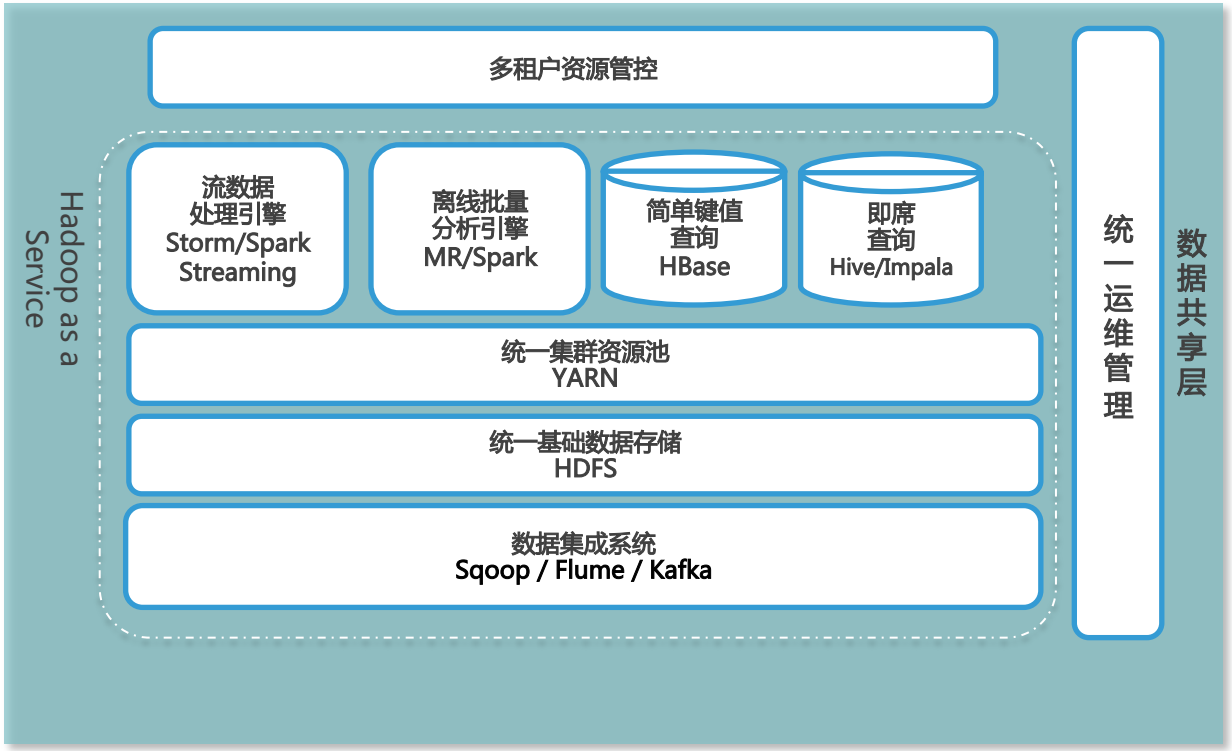
- Hadoop生态圈的优势
  - 开源的技术框架**大大降低了大数据处理的门槛和成本**，提供简洁有效的分布式计算存储的实现方式。
  - 生态圈日益完备，渐成**业内标准**

核心技术点：

技术理解 & 规划：Hadoop生态日趋复杂，需要业务和技术结合进行合理的架构和资源规划。

统一运维管理：开源架构、建设成本低，但技术支持不完备、更需要专业运维，同时持续跟进开源社区。

多租户资源管控：同一Hadoop集群需要支持不同业务需求，需要做好资源管控、任务优先级管理以及数据的隔离和共享



统一数据管理平台

Unified Data Management Platform (UDMP)

数据源

第一方数据

线下CRM、企业网站
移动应用、电商.....

第二方数据

微博/微信粉丝/线上营销搜索广告展示广告...

第三方数据

QQ、微博
豆瓣、论坛...

平台对接

数据融合

精准
用户画像
+
统一
ID体系
(手机号码、
邮箱、社交号)

用户属性

用户行为

用户偏好

用户价值

数据应用

消费者洞察
产品口碑

广告
精准投放

个性化
搜索/推荐

第三方接入

数据体系规划 & 平台构建

互动派

应用数据层

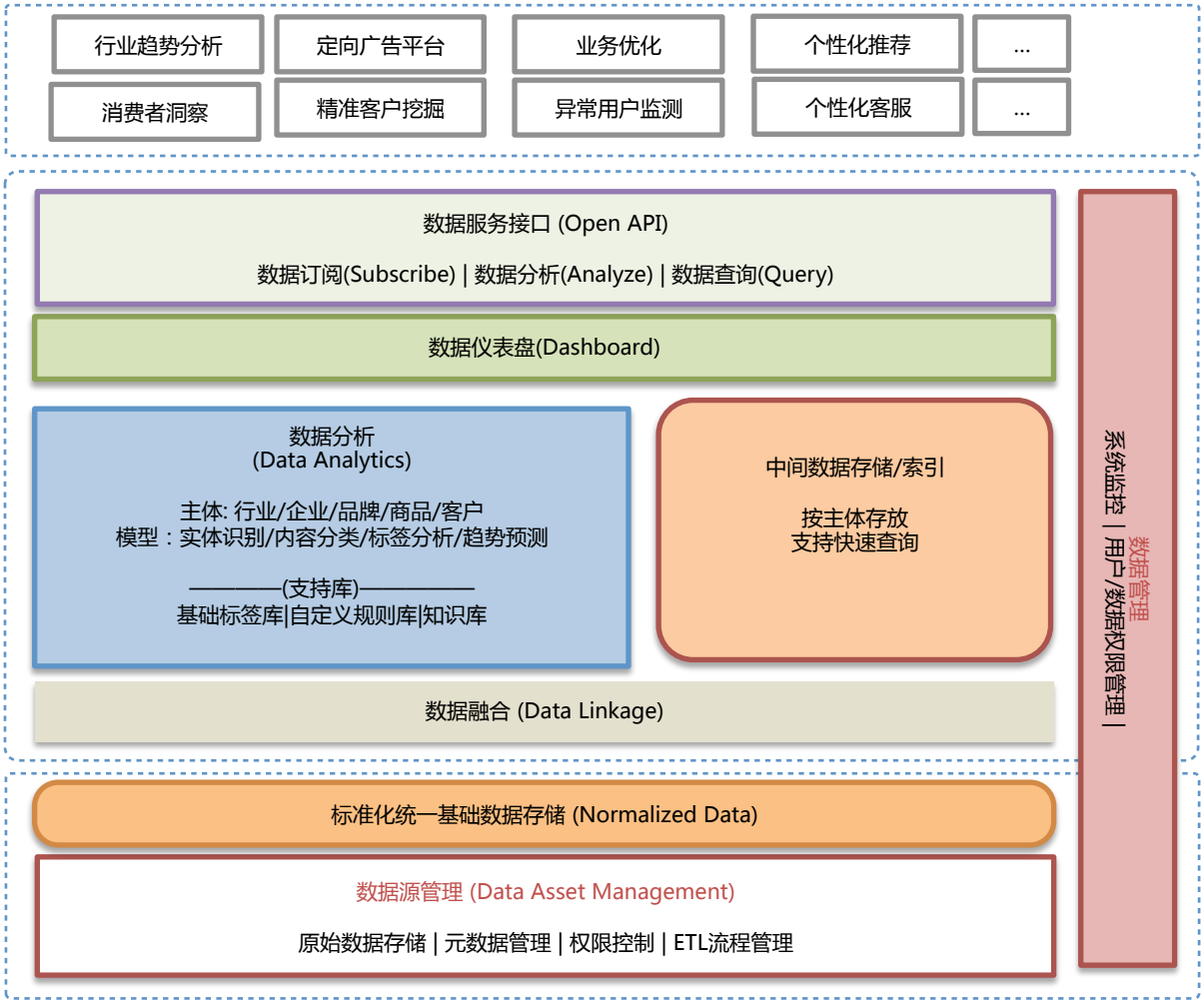
针对特定应用和问题的数据
按应用存放

中间数据层

可再生数据资源，共性数据需求
基础层之上加工的数据，根据不同业务需求，按不同主体存放
通过数据管理平台，提供给应用使用

基础数据层

不可再生数据资源
原始数据，不做汇总，应用不直接访问，仅做简单ETL和数据清洗



企业大数据核心技术要素

如何获得丰富的数据资源？	互联网数据采集
如何处理分散而复杂的海量数据？	海量作业平台
如何深度分析挖掘让数据具有价值？	语义分析
如何以用户为中心构建数据管理平台？	用户画像

互联网数据采集

核心技术点

- 多租户/多实例的使用模式; 自主开发，基于Spark进行分布式批量/实时数据抓取
- 爬虫任务请求级别最小力度拆分，数据DAG，防止单个环节瓶颈
- 基于yarn的网络IO隔离，完善的告警、监控
- 支持各种反爬策略
  - 代理IP服务，统一调度代理IP使用
  - 移动端token获取
  - 自动化的反爬探测，选择合理的反爬规则

海量计算

核心技术要点

- 基于计算平台之上 JOB（任务)+TABLE(仓库) 的管理
- 丰富的内置任务类型
- 完善工作流调度以及失败重试
- 支持自定义任务满足企业需求
- 良好的结合yarn队列，对多租户进行管理
- 配套提供完善任务监控机制、使用统计

应用场景

多租户、复杂数据处理流程，单一框架无法满足需求，大量数据处理任务需要管理协调。

应用案例——

- H5数据监测: 每天亿级H5页面PV的分析处理、关系传播图绘制、KOL的发现、用户画像
- **社交媒体数据分析应用: 亿级用户数据分析，百亿级关系图，日处理百万级微博用户数据更新**
- 运营商DPI数据应用: 每天百亿级移动用户访问记录的分析和建模

语义分析

核心技术

覆盖语义分析主要基础环节

→词级分析

词向量模型(Word2Vec)

词分类/实体识别/词义消歧/关键词抽取

→句子级分析

句子向量模型(ParagraphVec) /句法分析模型

短文本聚类/情感分析/观点抽取

→文章级分析

内容分类/内容聚类

(单文本/多文本)内容摘要

技术要点

长期知识库/语料的积累

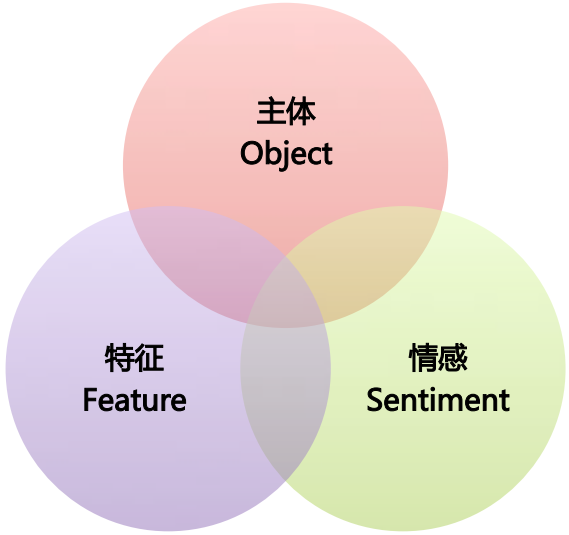
基于深度学习的核心语义模型（词/句子向量模型）

应用场景

各种非结构化数据分析

语义分析

从任意非结构化自然语言中，提取句子最精要的观点，即OFS三元组。



今天在山姆店买披萨，服务员(Object)态度(Feature)很不好(Sentiment)

附近就这一家沃尔玛(Object)，东西(Feature)比较齐全(Sentiment)

传统抽取方法 vs 三元组技术

传统抽取：酒店评价的简单关键词

传统抽取：中文歧义普遍存在

苹果 = 知名品牌？一种水果

沃尔玛服务员大大的赞
= 服务员大？服务员-赞

声音大 = 正面评价？负面评价

传统抽取：词汇受限于人的经验

厕所、洗手间、卫生间、男厕、女厕

还有什么词？

三元组：主体-特征-情感 的观点搭配

正面

负面

三元组：解决实体、特征、情感歧义

上下文词向量模型，消除实体歧义

语法分析、相关性分类器，确定特征搭配

极性词典，解决正负面歧义

三元组：词向量模型自动补全近义词

厕所间、蹲坑、蹲位、女厕所、男厕所、公共厕所、公厕、洗手池、浴室、洗澡间、马桶、洗漱间、洗手盆、内急、抽水马桶、洗手台、洗澡房、小解、洗手盘、蹲厕所、茅房、更衣间、解手、盥洗室、便池、小便池、wc、如厕

（说明：因为评价主体都是酒店，所以词云图中为“特征-情感”二元组）

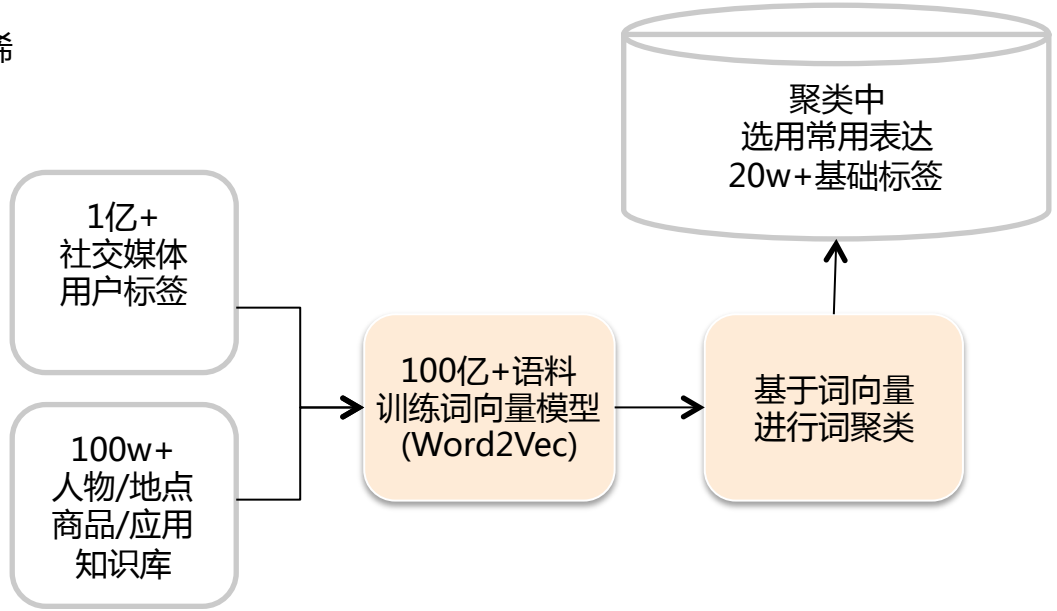
用户画像: 标签体系

传统方法问题

- 固定的标签集合很难满足业务的弹性
- 上层标签太通用，没法描述特定细分人群
- 简单关键词模型对通用类别词效果不佳（非常用表达）
- 高质量多层多分类模型训练数据稀缺，类别太多人工标注也很困难

标签体系要求

- 要有足够的覆盖面
- 要有足够的表达能力
- 要有足够细粒度
- 要能理解其背后的语义



用户画像: 标签模型

个人属性
年龄/性别， 行业/职业， 人生状态/个性
固定维度, 类别较少， 上层概念，独立模型

行为属性
网站/APP/新闻/视频

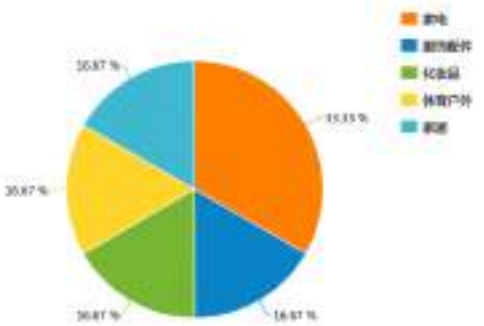
位置属性
行为轨迹 + 地点

兴趣属性
商品/品牌/美食/影视/图书, 维度巨大，有层次结构，
实体识别/知识库支持

关系属性
和谁交往 / 关注谁
关系分析

购物属性
商品库 + 商品识别/分类

价值属性
购买力/影响力/活跃度
构建特定指标及对应分类模型



04 大数据应用案例

大数据应用：数据驱动的零售渠道管理优化

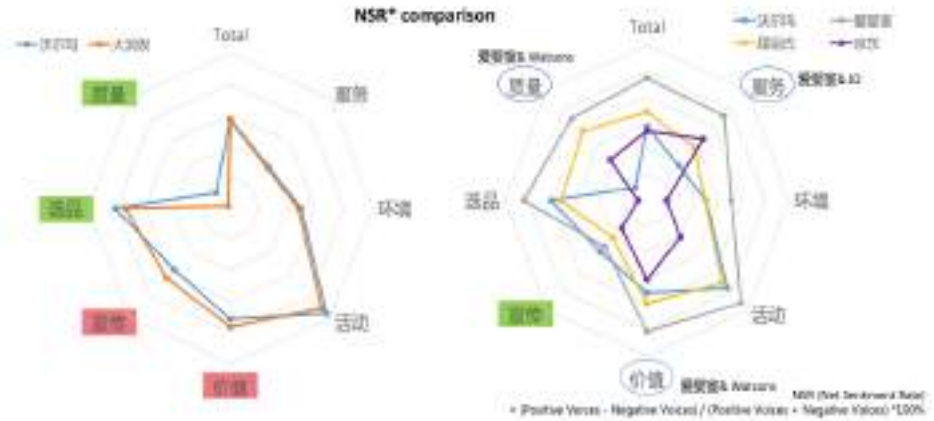
利用互联网用户评论数据进行社群聆听，监控与该日化企业合作的50个零售商店相关的用户评论，**通过线上数据进行渠道/购物者研究并指导渠道管理优化。**

实现过程：

1. 锁定微博、大众点评等互联网数据源，采集百万级别消费者谈及的与该日化品牌购物相关内容；
2. 利用自然语言处理技术，对用户评论进行多维建模，包括购物环境、服务、价值等10多个一级维度和50个二级维度，实现对用户评论的量化；
3. 对沃尔玛、屈臣氏、京东等50个零售渠道进行持续监控，结果通过DashBoard和周期性分析报告呈现。

意义：

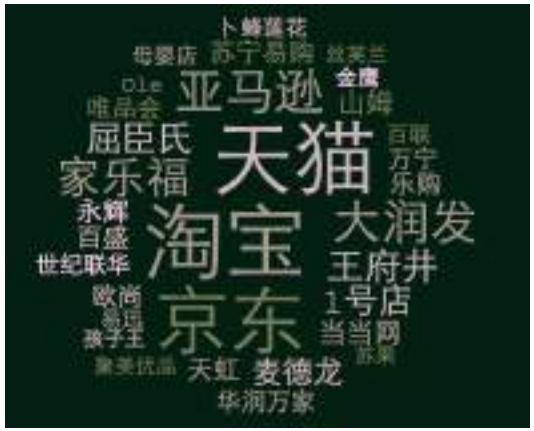
1. 有效掌握KA渠道整体情况：
   - 渠道：大卖场超市、综合商场、高端超市、化妆品店、母婴店、
   - 城市：一线、二线、三线
2. 了解KA渠道的关键细节、优势与劣势
   - 环境：货架、布局、设施、空间、位置、人气
   - 服务：态度、效率、专业性
3. 指导渠道评级体系调整和产品促销规划
   - 价值：价格、促销、会员、比价
   - 选品：齐全性、独特性、进口、新品

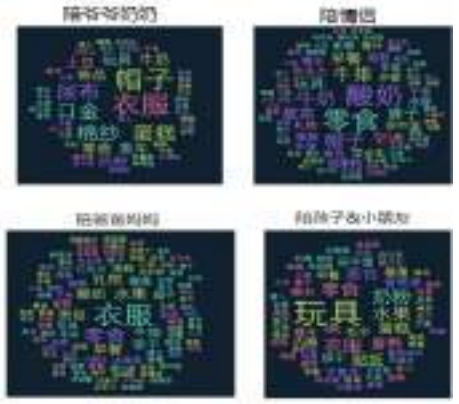


大数据应用：数据驱动的零售渠道管理优化

对购物者评价进行语义挖掘，实现零售渠道的店内核查，分析消费者购物习惯

价值	宣传	服务	活动	环境	质量	选品	支付	网店体验	商品类型	执行	产品品牌	产品类型	陪伴者	购物目的
价格	宣传	业务	活动	人气	质量	独特性	支付安全	功能	品牌	价签	化妆品	化妆品	陪同事	急用
会员		主管		位置		齐全性	支付方式	性能	新品	质量	家电	家居	陪同学	补货
促销		后勤		货架				移动端	进口		数码	家电	陪外公	采购
比价		售后		布局				设计			母婴	数码	陪外婆	闲逛
		接待		空间				其它			食品	服装鞋靴	陪奶奶	
		收银		设施								母婴	陪妈妈	
		送货		其它								生鲜	陪孩子	
		其它										箱包饰品	陪室友	
												食品	...	





大数据应用: 数据驱动市场和消费者研究

行业

行业现状及趋势洞察

大数据雷达对乳制品行业生命周期及行业销量进行趋势分析，评估行业生命力与健康度；分析竞品市场占有率及与竞品区域市场比重，以寻求市场突破口或抢先产品、渠道等市场布局。

品牌

品牌健康度研究

通过大数据雷达系统从其品牌定位、品牌认知、品牌形象、品牌美誉、舆情预警、品牌资产等维度分析消费者对该乳制品品牌的认知匹配度、品牌喜好度、评价关键词正负比例、情感卷入度等，并与竞品进行对比分析，以综合评估品牌健康度。

产品

产品分析

大数据雷达系统对产品销售趋势、市场份额比重分析，判断产品所处生命周期阶段；分析消费者对产品功能、性能、包装、质量的市场关注度、认可度，为新需求发现、产品研发、定价、渠道布局、营销推广等提供策略指导。

消费者

消费者画像与需求洞察

通过大数据雷达系统分析人口属性、关系属性、内容属性；发现其潜在兴趣点和消费行为特征，以为该乳制品在营销策划、内容创作、日常运营等提供策略和个性化服务需求。

媒介

营销活动媒介策略规划

大数据雷达系统对投放媒介实行前期监测、挖掘传播KOL，并为媒介组合提供精准策略，同时根据对投放媒介进行实时跟踪，优化媒介策略和效果评估。

渠道

重客画像与监测

通过大数据雷达系统分析重复消费者线上及门店购物偏好、购物行为、购物体验、评价情感值、服务体验、购买单价等数据，按人口属性、关系属性、价值属性、经济属性等维度进行标签画像，同时实时监测个体消费者的社交行为与购买路径，以更新用户数据，丰富重客标签。

大数据应用: 数据驱动市场和消费者研究

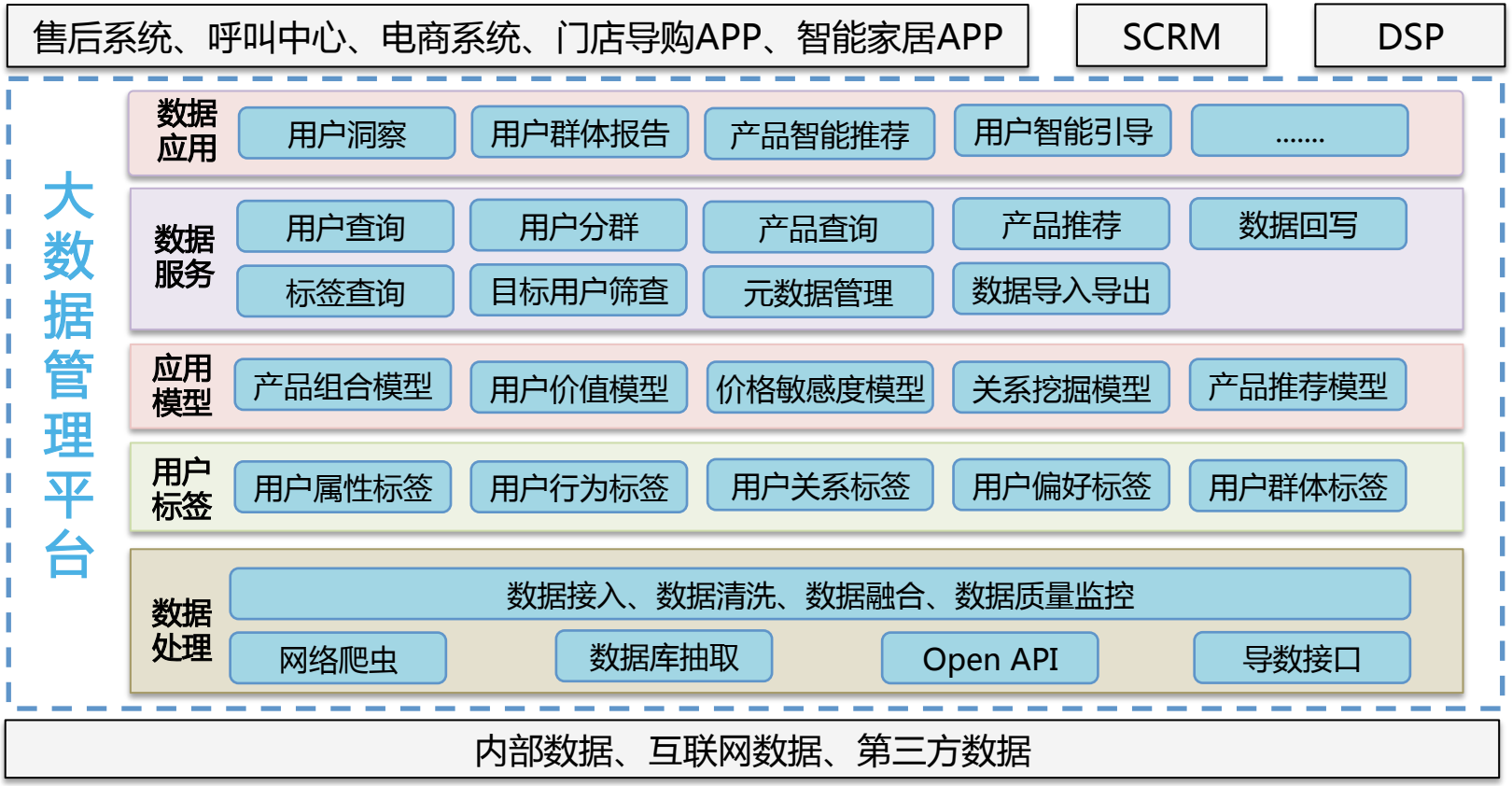
产品多维建模、产品优缺点分析、产品关注属性分析、竞品对比

产品多维度分析

产品取胜点在哪？

用户最关注的产品属性是哪些？

某家电品牌大数据管理平台：应用架构



建设统一数据管理平台，实现用户全景视图并构建业务支撑应用
将沉淀的数据资产转化为生产力

某家电品牌大数据管理平台：数据架构